Research on Timbre and Musical Contexts at CCRMA

John Strawn
Center for Computer Research in Music and Acoustics (CCRMA)
Department of Music
Stanford University
Stanford California 94305
U. S. A.

Presented at the
International Computer Music Conference (ICMC)
Venice, Italy
September 27 – October 1, 1982

Introduction

During the past few years, a number of advances have been made in the fields of signal processing, artificial intelligence, and computer science which now permit us to extend timbre research in a number of ways heretofore difficult if not impossible. This paper reviews our current efforts to investigate the nature of timbre and timbre perception, especially in situations involving more than one isolated note.

1. What is the "Timbre" of an Individual Musical Tone?

"Wer wagt hier Theorie zu fordern!"[1]

No one seems to be able to define "timbre" in a generalized and satisfactory way. The definitions which *have* been used to date have delineated

---

Given at the session on Psychoacoustics, 30 September 1982. Written version of 19 October, 1982. To appear in the conference *Proceedings*, published by the Computer Music Association, P. O. Box 1634, San Francisco, California, 94101. Typeset with TEX, the document compiler created by Don Knuth.

1. "Who dares to require a theory for this!" — The final sentence of Schönberg's *Harmonielehre* (my translation).

and in some ways restricted the range of questions being asked about timbre perception. In this section, I would like to explore whether these limitations can now reasonably be relaxed.

1.1. Definitions of "Timbre"

The official English-language definition of timbre for individual tones runs as follows:

> "Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar. ... Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus" (American Standard Acoustical Terminology, 1960).

This definition has been cited, even if not explicitly followed, in the research reported by Plomp, Grey, Risset, Wessel, and others.

A wide variety of popular definitions is also available. The *Petit Larousse illustré* has "The quality which differentiates two sounds at the same pitch and the same intensity,"[2] which echoes the opening of the ASA definition. A German "dictionary definition" of *Klangfarbe* seems on the surface to be more "scientifically" based: "the peculiar characteristic of a tone which is determined by various arrangements of the overtone series."[3] The *Harvard Dictionary* harks back to the opening of the ASA definition: "The quality ('color') of a tone as produced on a specific instrument, as distinct from the different quality of the same tone if played on some other instrument" (entry for "tone color", p. 856).

To summarize, these definitions are based for the most part on 1) criteria of discriminibility of isolated tones, and/or 2) models (even if outdated) of perception. These definitions, and recent work on timbre, focus both on the issue of the perception of ***sound quality*** and the question of ***identifying*** some (musical) instrument (see also Grey 1975, p. 1, second paragraph).

---

2. "Qualité qui distingue deux sons de même hauteur et de méme intensité" (p. 1021).

3. "durch die verschieden gestaltete Obertonreihe bedingte Eigentümlichkeit eines Tones." *Der Sprach-Brockhaus*, p. 340. This wording, especially "Eigentümlichkeit," follows von Helmholtz' useage closely (1877, p. 118).

1.2. What Does "Timbre" Include?

At an early stage of research in auditory perception, "timbre" was separated from pitch and loudness. If the sensation of pitch changes with the period of the waveform, and the loudness changes with amplitude, then timbre will change, so went the reasoning, if the shape of the waveform changes. In light of this reasoning, it is easy to see why the question of phase seemed so important to researchers, especially in the 19th century: when phase changes, the waveform changes drastically. It is now commonly accepted (Plomp, 1976) that phase plays a very minor rôle in timbre perception.

Thus, something besides pitch, loudness (to follow the ASA definition), and waveshape allows us to tell two tones apart.

Von Helmholtz recognized that the attack or the decay can affect the percept.[4] It is possible to perceptibly change the attack within limits, leaving the "steady-state" alone, and the listener can still correctly identify the instrument. (Errors in identification may increase.) But when we change the attack, so that the change is noticeable — have we then changed the timbre? Or have we simply changed the attack? The same holds, of course, for decay.

It seems commonly accepted that timbre does not include spatial cues (size of hall, location of source, motion of source relative to listener); indeed, Grey (1975, p. 1) ruled out "spatial location" explicitly. If we work with the ASA definition, this seems reasonable, since that definition requires that two tones be "similarly presented." However, in a given hall, it should be possible to distinguish between recordings of the same acoustic event made in two different places (see Plomp and Steeneken, 1973). Do these two recordings have the same timbre?

There are other aspects of musical sound which are difficult to fit into the definitions of timbre cited above. Some of these will be discussed in Section 2.4.1. If we add vibrato to a vibrato-less tone, for example, or change the amount or rate of vibrato — have we merely changed the vibrato, or have we modified the timbre?

The ASA definition quoted earlier does not mention duration. Grey again took great care to equalize the tones compared in his experiments so that they all had the same duration (as well as amplitude and pitch); von Bismarck (1974a) did the same. If the duration of an event is changed, is the timbre changed? At least in extreme cases, the vagaries of our perceptual system may help bring about a change in timbre. Risset reports that " ... details

---

4. " ... alle möglichen verschiedenen Eigenthümlichkeiten der Klänge ... von der Art und Weise abhängen, wie die Klänge anfangen und enden ... " (von Helmholtz, 1877, p. 114.) "... all possible peculiarities of musical tones ... depend upon the way in which they begin and end." Ellis' translation, p. 66.

of the attack [of trumpet tones] were more important aurally [i. e. audible] for a long and sustained tone than for a short one — as though, in the latter case, the ear were immediately 'distracted' from the attack by another nonstationary event" (1966, p. 36).

This line of reasoning suggests that all of these properties of the physical signal may influence the perception of timbre. Indeed, Grey points out (1975, p. 1) that "by timbre, the musician indicates the tonal qualities which characterize a particular musical sound." If one were to follow this definition, it would seem reasonable to ignore the restrictions which the ASA definition places on pitch and loudness; after all, in some cases, merely changing the *loudness* of a musical tone may evoke a different timbre percept in the musical listener. Schönberg apparently felt the same way about *pitch* and timbre (Carter translation, p. 421):[5]

> The distinction between tone color and pitch, as it is usually expressed, I cannot accept without reservations. I think the tone becomes perceptible by virtue of tone color, of which one dimension is pitch. Tone color is, thus, the main topic, pitch a subdivision. Pitch is nothing else but tone color measured in one direction.

(This sentence, on the next-to-last page of *Harmonielehre*, leads into the discussion of *Klangfarbenmelodie*).

1.3. What is a Steady-State Tone?

In timbre research, there are historical precedents for limiting the kinds of signals studied in order to simplify matters and to avoid the anarchic state of affairs which the previous section managed to reach.

As mentioned at the beginning of the previous section, there was a tendency to equate the shape of the waveform with the timbral percept evoked by the waveform. A mathematical technique for analyzing this waveshape was conveniently provided by Fourier's theorem. But the theorem requires that the waveform being analyzed not vary in time. This means that attacks and decays cannot (strictly speaking) be analyzed with Fourier's methods. Von Helmholtz, for this and perhaps other reasons, restricted his

---

5. Ich kann den Unterschied zwischen Klangfarbe und Klanghöhe, wie er gewöhnlich ausgedrückt wird, nicht so unbedingt zugeben. Ich finde, der Ton macht sich bemerkbar durch die Klangfarbe, deren eine Dimension die Klanghöhe ist. Die Klangfarbe ist also das große Gebiet, ein Bezirk davon die Klanghöhe. Die Klanghöhe ist nichts anderes als Klangfarbe, gemessen in einer Richtung. (1922, p. 503).

timbre research to steady-state tones,[6] and called the percept evoked by such a tone "musical timbre."[7]

The perception of "steady-state" spectra has recently been extensively studied. Von Bismarck (1974a, b) created 35 different signals, some based on filtered noise, some based on (filtered) spectra of periodic tones. To measure these signals, he used 30 verbal scales such as "smooth–rough" or "compact–scattered." Test subjects were asked to rate each of the test tones on each of the scales. Factor analysis of the resulting data showed that four factors could account for more than 85% of the variance. The most important of these four factors seemed to be related to "brightness" (1974b). The importance of "brightness" had been suggested by other studies as well.

Von Bismarck's conclusions about his data seem perfectly reasonable. However, this research, like most research with "steady-state" signals, avoids one crucial issue. Von Bismarck writes that "[i]n order to avoid audible on- and offset transients, the tape endings were cut with an angle of 45° which led to on- and offset times of approximately 16 msec" (1974a, p. 151). The total duration of each stimulus was 5 seconds.

I submit that these are not steady-state tones. Instead, they are very long tones with *one type of attack and decay.* It would be interesting to see if data such as collected by von Bismarck would be influenced by using other attacks and decays. Performing musicians are used to applying verbal attributes to attacks and decays: a conductor tells the performers, "give me a sharper attack," or "don't put so much bite into the attack;" and the like. If the percept evoked by an attack can influence the timbre of the overall event, then it is difficult to generalize von Bismarck's verbal attributes beyond this particular set of attacks, decays, and signal durations.

As for research in musical contexts, recent work (especially Grey 1975) has shown that there is no reason to be restricted to steady-state tones.

2. Timbre Perception in Musical Contexts

The first section of this paper discussed problems with the historical and current definitions of timbre. This section will show how research in musical contexts may contribute to a better understanding of the nature of timbre itself.

Perception of music in musical contexts is a vast field; much of the work which has been done will not even be mentioned here. I am investigating just

---

6. " . . . wir ... berücksichtigen nur die Eigenthümlichkeiten des gleichmässig andauernden Klanges" (1877, p. 116). " we shall . . . confine our attention to the peculiarities of the musical tone which continues uniformly." (Ellis translation, p. 67).

7. "Musikalische Klangfarbe" (1877, p. 118). This term seems to have died out in German useage. I have found it in Backhaus (1932) but not (yet) in Winckel (1960).

one part of the field. Speaking in professional musical terms, even an isolated note is played in a "musical context": surrounded by silence. One can also consider "musical context" as the hall in which the performance occurs, as the position in the current movement/piece, etc. Unless otherwise stated, I will use "musical context" to mean "a note surrounded by other notes." In other words, I refer to "musical context" in order to contrast my work with research historically conducted using isolated tones. "Surrounded" can stretch into the "horizontal" or "vertical" direction. I am interested, then, in examining the perception of musical timbre and how it is affected by playing not just one note, but several, either together in polyphony, or as a melody.

2.1. Modelling Musical Contexts

Grey (1975) and others conducted extensive research into the perception of individual music notes. Our initial approach is to view, say, the musical context of a melody as a chain of notes. An initial question is: how does the note played in the musical context differ from the note played by itself?

As a first model, I would suggest the following: the individual amplitude and frequency functions representing the harmonics of a note may be changed in characteristic ways, depending on such factors as the intended articulation, the place of the note in the phrase, and so on. Max Mathews, for example, in examining slur formation, has had some success in creating phrase-like percepts by simply overlapping amplitude envelopes of two successive notes and adjusting the overlap times (Mathews, 1982). Dexter Morrill (1980) has conducted an extensive study of "phrase information," and found, among other things, a characteristic "phrase envelope," an amplitude envelope which modifies the amplitude envelopes of the individual notes in a phrase.

However, such attributes of phrases are not the focus of our work. Rather, I am interested in the mechanism of timbre perception, and especially in finding out what contributes to the timbre percept of the individual notes in musical contexts. This model already suggests avenues of research: for example, in a simple two-note passage, if we leave out some of the attack information necessary when the second note is played by itself, will the perceived timbre of the second note be changed? Indeed, at the level of the signal, what happens between two notes for various kinds of articulation? Are the frequency functions for two successive *legato* notes connected? If so, how?

2.2. Source Discrimination

In polyphonic contexts, the question arises: how does the "ear" sort out the individual notes? Without going into a great amount of detail here, I will adopt McAdams' interpretation (1982) of experimental data. According

to this view, the individual spectral components impinging upon the ear are first separated by the perceptual system into one or more of what I will call "virtual sources." I would extend this notion to say that timbres are assigned thereafter to the virtual sources (see Figure 1). Likewise, pitch, loudness, location, and other perceptual attributes of sound are assigned *after* the spectral components have been assigned to virtual sources. Although this idea may seem obvious at first, it has various important implications, a discussion of which will occupy much of the rest of this paper.

2.3. Timbral Quality versus Instrumental Identification

It seems to be common useage that if we filter a given sound (even a "steady-state" artificial stimulus), so that there is a perceptible difference in the sound, we say that we change that sound's timbre. (By filtering, I mean here linear filtering, i. e. changing the amplitudes and/or phases of the spectral components. I am not interested here in the changes in loudness which the filtering may evoke). In other words, the changes brought about by filtering are generally described as a change in timbre. In many cases, for musical instruments, changing the timbre in this fashion will not prevent us from identifying the instrument (see Risset, 1966, p. 12). Consider the poor transmission characteristics of the filter given by car radios and speakers. In spite of the spectral changes which such equipment presumably produces (including missing fundamentals due to the small speaker size), I have been able to correctly identify an English horn (as opposed to an oboe) on my car radio.

Based on considerations such as these, I believe that the *perception of timbral quality* can be considered as a different process from the *identification* of a traditional musical instrument.[8] In the next three sections, I would like to discuss some differences between these two.

2.3.1. Timbral Quality

The percept of "timbral quality" seems to have access to certain innate associations. Many researchers (e. g. von Helmholtz) have devoted considerable effort to providing everyday labels to a wide variety of stimuli. Indeed, the mere fact that von Bismarck's test subjects *could* rate his artificial stimuli on a wide variety of verbal scales without large variations in the ratings (von Bismarck, 1974a, Table I) suggests that there are some innate as-

8. Von Bismarck's instructions to his test subjects included: "Do not attempt to identify the source of a sound . . ." (1974a, p. 158).

sociations evoked by timbral quality. Other perceptual systems (e. g. vision) seem to have innate notions (like "surface") as well (Bregman 1977, p. 253).

2.3.2. Instrumental Identification

The ability to identify instruments, on the other hand, is probably "acquired through a learning process" (Risset and Wessel, 1982, mss. p. 24). Perception of timbral quality differs from identification of instrumental sources in that quality can change drastically while the listener knows that the source has remained the same. This happens, for example, when brass or strings play with mutes.

It has become convenient to discuss timbre perception in terms of some n–dimensional perceptual and/or stimulus space (Grey, 1975; Plomp, 1976). Presumably, any possible musical tone can be represented as a point in such a space. If this is a perceptual space, then explicitly identifying timbral quality can be likened to identifying the coordinates of a sound in that space. As for identification of instrumental sources, McAdams points out that "[e]ach instrument would have a bounded region [in the space], though there is certainly the possibility that different instruments' regions would overlap. Those instruments with a wider pitch range and more timbral versatility would occupy larger or even multiple regions" (1982, p. 295). Identifying the instrument would thus correspond to selecting which one(s) of the "instrumental regions" contain the given point. But I would warn against oversimplifying here; McAdams is correct is pointing out that a given instrument can occupy multiple regions. That squawk produced by a beginning clarinet player can still be identified accurately as coming from a clarinet! In these terms, *connecting* these separate regions and attributing them all to a single source would seem to be a learning process.

2.3.3. Interdependence with Music Perception

Let us return for a moment to the model that says that timbre is assigned *after* the partials impinging upon the ear have been assigned to one virtual source. McAdams suggests (1982, p. 289) that "the processes that synthesize source images are continually comparing the 'current' image(s) with the new information from the environment and when enough information has been acquired to indicate that this interpretation is no longer valid, a new synthesis is effected and new source image(s) formed." I would like to suggest, in turn, that timbre quality or even the instrumental source ***need not be explicitly assigned for every note in a musical context.***

Consider the following musical context: a trombonist in a Dixieland jazz band. In a fast passage, such as shown in Figure 2, a trombonist might be playing riffs consisting of relatively long notes interspersed with passages of several vivace notes ("A" in Figure 2); or very quick notes like grace notes

might occur ("B" in the figure). Does the "ear" explicitly assign a timbre to each of those fast notes? I don't think so.

To generalize from this assertion, it seems reasonable to assume that initial, explicit timbre assignments are quickly made in (polyphonic) musical contexts, such as the first note or so in Figure 2. If the listener is attending to other aspects of music, such as pitch, perhaps the explicit timbre assignment is put off for a short while. Also, visual information may make explicit timbre assignment by the auditory system unnecessary. Once timbres have been assigned, then the individual musical lines are followed, and timbre is not explicitly re-assigned until there is a reason to do so, for spectral or musical reasons. This might happen at "A" in Figure 2. In other words, the mechanisms which "track" melody, voices in polyphony, harmony, rhythm, etc. can provide enough information to the timbre classifier so that its work is unnecessary; in many cases, timbre will not be explicitly processed.

I would further suggest that as the music is being played, there are certain predominant notes at which the current timbre assignment is re-confirmed, in addition to whatever visual cues may be available. Such notes would stand out musically and/or perceptually from the musical context: loud notes, peak notes in a melodic phrase, notes accented in some way.

2.4. Some Aspects of Timbre Perception

Having separated the notion of musical quality from that of instrumental identification, I would now like to turn to a few selected aspects of timbre perception.

2.4.1. The Rôle of Cues

Grey (1975) found that minute details in the amplitude and frequency functions for additive synthesis can be left out with little or no change in the percept. Larger-scale time-varying changes have only recently been studied extensively (but see Backhaus, 1932). These time-varying features are on the order of, say, tens of milliseconds in the "attack" and perhaps slightly longer in the "steady-state" — such as the "blips" in the attack function which seem to be characteristic for brass tones.

There are also larger-scale characteristics of sound which are important for timbre perception in music. An example of this is given by the manner in which the spectral components sweep slightly through the formants of a stringed instrument when the player is using vibrato.

I would like to subsume these kinds of phenomena under the term *cue*, and suggest that (time-varying) cues provide a basis for *instrument recognition.* Indeed, gross spectral characteristics, such as the manner in which the even-numbered harmonics tend to be damped in the clarinet, may be included as cues.

Not only the existence of such cues, but also the manner in which they change together, may be important for timbre perception. This notion is inspired by Bregman's work (1977) with perception and behavior. In Bregman's terms, an "ideal" (here, a cue) is "instantiated" in a percept, or in behavior. "Ideals are altered as they enter into composition with one another; these alterations will be called 'transformation' " (Bregman 1977, p. 254). as one aspect of the percept, such as pitch, changes due to such a compositional process, then other aspects (here, attack characteristics, spectral envelope, etc.) change in a characteristic, concomitant way.

The multi-dimensional perceptual/spectral spaces constructed to date (Grey, 1975; Plomp, 1976) do not take seem to take such cues as these into account. In fact, the spectral dimensions contributed by these cues may be orthogonal to the axes in these multi-dimensional spaces. This could bring us to yet another anarchic state of affairs, in which any isolatable cue would contribute another dimension to the space. Liberman has indeed remarked that in examining the relationship between acoustic cues and the perception of speech, "[c]ertainly every potential cue so far tested has proved to be an actual cue, no matter how peculiar seeming its relation to the phonetic segment" (1982, p. 151). One way out of this dilemma is to accept the notion that some cues may be more important than others both in the perception of timbral quality and in instrument identification (see Liberman, 1982, p. 151; Risset and Wessel, 1982, mss. p. 21).[9] This is reasonable, given Risset's success (1966) in brass tone synthesis based on just a few features. This still does not answer the question of what rôle these cues play in timbre perception.

Such cues may play an important rôle in timbre interpolation and categorical perception, which Grey (1975) and Wessel (1979) studied. Consider the case of interpolating from a brass sound to a clarinet sound. Assume further that the "blip" in the brass attack is important for identifying the brass instrument, and that such a blip is missing in the clarinet attack. Likewise, assume that the steady-state trumpet spectrum lacks the characteristic weakening of the even-numbered harmonics found in the clarinet spectrum. Grey (1975) was able to demonstrate that categorical perception of timbre does not occur for a "linear" interpolation between these two "endpoint" timbres. As Grey himself warns (1975, p. 76), his conclusion is valid only for the path which his interpolation scheme followed through the multidimensional timbral space. It might be possible to show that categorical perception does in fact occur if more than just time-varying amplitudes and frequencies of the spectral components provide the basis for interpolation.

---

9. In Bregman's terms, "As the pressure of an ideal becomes stronger, it becomes more visible to other ideals" (1977, p. 280) in the perceptual composition of ideals.

2.4.2. Data Reduction and the Rôle of Critical Bands

It has been suggested recently that *critical band* phenomena may play a rôle in timbre perception. Briefly, a critical band is a region encompassing approximately a third of an octave. Certain aspects of hearing can vary depending on whether the spectral components (or bands of noise) being tested fall within one critical band. This is true, for example, for aspects of the perception of loudness, phase differences, and some forms of masking. Even listener's ability to "hear out" partials of a tone can be related to critical bandwidths (Plomp 1976). It should be emphasized that critical bands do not form a fixed set of filters, although computer-based analysis/synthesis methods based on such a fixed set of filters have proven useful for some purposes (Petersen 1980).

Since data reduction for additive synthesis of isolated instrumental tones (Grey, 1975; Charbonneau, 1981) has been so successful, further data reduction might be possible in musical contexts without altering the timbre percept. Initial work (Grey, 1978) suggests that the ear's ability to detect fine temporal differences in tones is reduced in musical contexts, which would also make further data reduction seem possible.

Assume for a moment that several spectral components of a tone fall into one critical band. Assume further that only one of these components influences timbre, and the others are effectively ignored due to critical band phenomena. This would mean that we could reduce all of these components to one. To test these assumptions, we could re-synthesize tones based on the data-reduced spectra, and listen for audible differences. If no differences were audible, then this would imply that the omitted data (based on critical bands) is not important for timbre perception.

However, further reflection on the perception of timbre in musical contexts suggests that such an approach can become too simplistic. In musical contexts, the ear can separate two or more notes played together, even those played in unison. To name only one case: consider the *continuo* bass line in baroque music; it is possible for the listener to separate the bassoon or cello following note-for-note the bass line played in left hand on the organ or harpsichord.

In such a case, and in polyphony in general, some of the spectral components from each tone will fall into one "critical band," especially in higher frequency ranges. One might assume that detailed information about these spectral components from the separate tones is lost in this case, and is therefore not important for timbre perception. But certain *time-varying* information, such as coordinated vibrato in all of the spectral components of one note, can and does "pass through" critical bands to provide the basis for identifying sources (McAdams, 1982). (I am assuming here that critical band phenomena occur before spectral components are grouped into virtual sources.) This implies that separate spectral components falling within one critical band can still be resolved by the ear for some purposes. It seems reasonable to assume that this time-varying information is also available for

instrument identification and assigning timbral quality. The same would hold for the time-varying information implied by "timbral cues" (see Section 2.4.1). Thus, critical band phenomena would not affect the importance of these time-varying cues for timbre perception.

One last point: Bregman has suggested (1982) that data reduction may not be possible to a large extent in musical contexts: it may be necessary to retain the time-varying information in the higher-order harmonics in order to allow the ear to parse sources adequately!

2.4.3. The Rôle of Reverberation

In this section I would like to temporarily expand the limited definition of "musical context" given in Section 2 and consider the question of the perception of musical timbre in performance. In most listening situations, one physical sound source results in a large number of reflected virtual sources which reach the ear, resulting in the mixed percept of the sound source plus reverberation. The phases and amplitudes of the spectral components are modified as these reflections are mixed. This raises the question of how reverberation affects timbre perception.

Experimental evidence shows the conditions under which a change in phase can affect timbre perception. Plomp (1976, p. 110) found that "the effect of phase on timbre is greatest if all harmonics of one tone are either sine or cosine terms and the harmonics of the other tone are alternate sine and cosine terms." Of course, such conditions are unlikely to occur in nature. He also showed that the change in timbre caused by this change in phase was small compared to the effect on timbre of changing amplitude.

In normal listening situations, reverberation scrambles phase so drastically that reverberation *alone* makes it unlikely that timbre is determined by waveshape or the phase relationships between spectral components. Reverberation also affects the relative amplitudes of spectral components, at least to some degree. These effects are compounded by the fact that the musical signal itself is constantly changing.

Research has only begun on the effects of reverberation on timbre perception (Plomp and Steeneken, 1973). Given the success in modelling reverberant spaces in the past few years (Moorer 1979), we are now in an excellent position to research this question.

2.5. Toward a Model of Timbre Perception in Musical Contexts

By way of summary: In musical contexts, spectral components are grouped by the auditory system into virtual sources at an early point in auditory processing. Thereafter, timbre, loudness, pitch, location (distance/-angle/velocity), and the like can optionally be explicitly assigned. At least in some musical contexts, timbre is not explicitly assigned to every note.

3. Modelling the Physical Signal

Given a model of timbre perception, another question is how to formulate a model of the signal impinging on the auditory system. For perceptual studies on the questions discussed here, the time-honored "additive synthesis" model still seems completely adequate. In this model, we take a "snapshot" of the signal at some point in time and analyze the signal into spectral components. Each such snapshot can be thought of as the *line spectrum* of the signal at that point in time. Resynthesizing the signal on the basis of this analysis data produces an exact copy of the original (within the limits of computational accuracy). For perceptual research, we can modify the spectral components in some way, and investigate the response of the perceptual system to the modified signal.

3.1. The Phase Vocoder

In recent years, the digital phase vocoder has become popular for analysis of musical tones. This popularity is due in part to the fact that the phase vocoder, like certain other analysis techniques, is able to analyse a musical signal in a manner which is intuitively appealing (see Figure 3). The phase vocoder itself is computationally efficient and very robust. John Gordon and I have recently implemented[10] and rigorously tested a new version of the phase vocoder (Gordon and Strawn, 1984). Where possible, we have used the routines distributed by the IEEE (IEEE Digital Signal Processing Committee, 1979).

3.1.1. Problems with the Phase Vocoder

But in using the phase vocoder to analyze musical passages, we quickly run into some problems. The analysis channels in the phase vocoder are "set up" at the beginning of the analysis run, based (typically) on the fundamental frequency of the note and on the sampling rate. If a note at another pitch is analyzed with this setup, some spectral components may fall "across" two channels.

Figure 4 shows a "spectrographic" plot of two clarinet notes played *legatissimo*. Figure 5 shows the same analysis of the same tones on a channel-by-channel basis. Channel no. 4 in Figure 5 starts with the third harmonic of the first tone at a fairly high amplitude, followed by the fourth harmonic at a much lower amplitude (about 25 dB). (The "fuzziness" in the plot of

10. Our implementation is in SAIL. Rob Gross at Berkeley has almost completed a parallel implementation in C.

the second note in this channel can be attributed to the low amplitude.) If the transition between notes turns out to be perceptually significant, then perhaps some detail in the transition between *harmonics of the same order* must be retained. In this case, the fourth harmonic for the first note is not in channel 4, but in channel 5!

The problem is even more pronounced in Figure 6. At first glance, it looks as though the first harmonic of the first note *splits* into the first and second harmonics of the second note. Closer examination of Figure 7 shows that this is not the case. Rather, the second harmonic of the first note is essentially missing (Figure 7, Channel 2); and the second harmonic of the second note is apparently "approached from below" in this analysis. Returning to harmonics 4 and 5: Figure 7 also shows that the fourth harmonic of the first note (in channel 5) approaches the fourth harmonic of the second note (in channel 4) from above, as one might expect. The details of this transition have meanwhile been "lost" as the components cross channels.

It is possible but in this case unappealing to combine adjacent channels to provide analysis data for one spectral component.[11] This is not attractive because again it is difficult to know how to track amplitude and frequency functions during the transition from one note to the next while combining channels. It may be adequate to "eyeball" the transitions between notes and to construct by hand sets of amplitude/frequency functions which model the transition adequately. This is a tedious proposition at best. Analogous problems occur with more than one instrument playing at once, as is already well known from work on automatic transcription (Moorer, 1975; Piszczalski et al., 1975, 1981; Foster et al., 1982).

By way of parenthesis: The object of the analysis/synthesis system is to adequately characterize the perceptible features of the signal. Although we have been very careful in implementing analysis and synthesis systems (see Gordon and Strawn, 1984), we are lucky here to be relieved of certain mathematical rigor. In establishing an analysis/synthesis system for perceptual research or compositional work, as long as *are able to resynthesize a perceptually identical signal*, we know that any mathematical rigor which we may be violating in order to arrive at the resynthesis can be ignored.

3.1.2. Other Solutions Based on Line Spectra

As was suggested earlier, it seems overly simplistic to adopt for our research the model of critical bands based on fixed filter locations and bandwidths. Thus, I will not be adopting the critical band transform (Petersen 1980) for this work.

---

11. Andy Moorer wrote out the equations for this in 1979.

We have seen that it is difficult for the phase vocoder to reconstruct the exact frequency and amplitude variations in a form convenient for us when the musical signal varies widely in frequency. It may be possible to construct a less rigorous analysis/synthesis system which will do so. It would be nice if we could vary the frequency and time resolution of the phase vocoder so that it could track transitions between notes. This may well be possible. The Wigner distribution also looks promising in this regard (Claasen, 1980).

3.2. Other Spectral Models

As discussed earlier, the phase vocoder is based upon a model of a signal which involves a number of spectral components to be additively combined.

Another popular model for sound analysis is based on a sound source followed by a (series of) filter(s). The techniques of modern spectral analysis analyze the signal in terms of models to derive a representation of the sound source and of the filter(s).

It turns out to be difficult to coerce this approach into modelling the kind of line spectra that we're dealing with in perceptual research. We have a signal whose spectral components will be varying widely in time and frequency. It is possible to model each spectral (line) component with two poles and two zeros. But what happens when the spectral components start wandering around in frequency? In order to obtain a pole-zero model in this case, we must place an upper limit on the analysis time used to computer this model. But this causes spectral lines to appear as wider spectral bandwidths. Thus, the pole-zero model tends to yield "wide-band resonances" in place of the lines which are intuitively so appealing. the pole/zero pair track the spectral line. This approach especially runs into difficulties when new spectral components "appear" (as seems to happen in Figure 6), or a spectral component already present disappears. Dealing with the behaviour of the poles and zeros in such cases can require a great deal of interpretation. For this reason, I currently do not anticipate using these models for my research.

3.3. Towards eMerge: A Spectral Editor for Research and Composition

The need for some sort of editor was already hinted at above, in the discussion of editing phase vocoder outputs for successive notes in a melodic phrase. Such an editor should incorporate the usual editing functions (creating, copying, deleting, modifying) that we've all learned to love and hate in other editors. We will need other features; an editor for our purposes should be able to generate line segment approximations to one or more functions, for example (Strawn 1980). Also, it would be useful to be able to listen to tones synthesized from the (perhaps edited) functions.

Such an editor does not have to be limited to applications in psychoacoustic research. From a compositional point of view, Objed implemented in Toronto provides inspiration (Buxton et al., 1982). Objed is designed to allow the composer to deal with timbre compositionally while minimizing the difficulties of dealing with timbre as an acoustical or perceptual can of worms. (In Objed's terms, our timbral objects will be limited to the additive synthesis model).

Such an editor has existed at CCRMA in the past. As part of his thesis work, John Grey (1975) prepared an editor called SYNTH, which was used by him and others (at CCRMA, notably Charbonneau (1981)). Based on our experience with this editor, I am now in the process of implementing a new editor to supercede SYNTH (Strawn 1982). The phase vocoder and certain other programs write their outputs into a file known generically as a "merge file" (Strawn 1979, pp. 4–5), which provides the basis for the name of the proposed new editor. Figures 4 and 6 were generated using the parts of the editor which are already functional. If, as seems likely, CCRMA acquires some work stations (Chowning et al., 1982), I obviously plan to transfer development of this editor to that environment.

4. Conclusion

In this paper, I have attempted to outline the current state of research in timbre at CCRMA. We are critically examining the concept of timbre and the nature of timbre perception. We are also in the process of assembling tools to allow us to exploit in research on timbre perception the advances in signal processing, computer science, artificial intelligence, and perception which have been made over the past few years.

Acknowledgments

Much of the material presented here was developed in the course of discussions with (alphabetically) John Chowning, Alexander Bregman, John Gordon, John Grey, James A. Moorer, Dexter Morrill, Jean-Claude Risset, Earl Schubert, and Julius O. Smith III. I would like to thank Earl Schubert, Bill Schottsteadt, Kip Sheeline, and Julius O. Smith III for their critical comments on this manuscript. I would also like to thank Emily Bernstein, who recorded the two-note clarinet phrase shown in the figures.

References

American Standard Acoustical Terminology S1.1–1960. New York: American Standards Association, Inc., 1960.

Backhaus, H. "Über die Bedeutung der Ausgleichsvorgänge in der Akustik." *Zeitschrift für technische Physik* 13(1):31–46, 1932. Translated as "On the Importance of Transients in Acoustics" by John M. Strawn, August 1982. Manuscript.

von Bismarck, G. "Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes." *Acustica* 30:146–159, 1974a.

von Bismarck, G. "Sharpness as an Attribute of the Timbre of Steady Sounds." *Acustica* 30:159–174, 1974b.

Bregman, Albert S. "Perception and Behavior as Compositions of Ideals." *Cognitive Psychology* 9:250–292, 1977.

Bregman, Albert S. Personal communication, 1982.

Buxton, W., S. Patel, W. Reeves, and R. Baecker. "Objed and the Design of Timbral Resources." *Computer Music Journal* 6(2):32–44, 1982.

Charbonneau, G. 1981. "Timbre and the Perceptual Effects of Three Types of Data Reduction." *Computer Music Journal* 5(2):10–19.

Chowning, John, Chris Chafe, John Gordon, Patte Wood. "Studio Report: The Stanford Center for Computer Research in Music and Acoustics." Presented at the 1982 International Computer Music Conference, Venice, Italy.

Claasen, T. A. C. M., and W. F. G. Mecklenbräuker. "The Wigner Distribution — A Tool for Time-Frequency Analysis."
Part I: Continuous-Time Signals. *Philips Journal of Research* 35:217–250, 1980.
Part II: Discrete-Time Signals. *Philips Journal of Research* 35:276–300, 1980.
Part III: Relations with Other Time-Frequency Signal Transformations. *Philips Journal of Research* 35:372–389, 1980.

Foster, S., W. A. Schloss, and A. J. Rockmore. "Toward an Intelligent Editor of Digital Audio: Signal Processing Methods." *Computer Music Journal* 6(1):42–51, 1982.

Grey, John M. "An Exploration of Musical Timbre." Ph. D. Dissertation, Dept. of Psychology, Stanford University. Department of Music Report STAN-M-2, 1975.

Grey, John M. "Timbre discrimination in musical patterns." *Journal of the Acoustical Society of America* 64(2):467–472, 1978.

Harvard Dictionary of Music. Cambridge, Massachusetts: Harvard University Press, 1972. Second Edition.

von Helmholtz, H. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Fourth Edition. Braunschweig: Vieweg, 1877. Translated as *On the Sensations of Tone as a Physiological Basis for the Theory of Music* by Alexander J. Ellis. New York: Dover Publications, 1954.

Digital Signal Processing Committe, IEEE Acoustics, Speech and Signal Processing Society. *Programs for Digital Signal Processing.* New York: IEEE Press, 1979.

Kay, S. M., and S. L. Marple, Jr. "Spectrum Analysis — A Modern Perspective." *Proceedings of the IEEE* 69(11):1380–1418, 1981.

Liberman, Alvin M. "On Finding that Speech is Special." *American Psychologist* 37(2):148–167, 1982.

Mathews, Max V., and Joan E. Miller. "How to Make a Slur." Murray Hill, New Jersey: Bell Laboratories. Typewritten mss., no date. Lecture, Center for Computer Research in Music and Acoustics, Stanford University, Stanford California, 7 July 1982.

McAdams, Stephen. "Spectral Fusion and the Creation of Auditory Images." in Manfred Clynes, ed. *Music, Mind, and Brain*: The Neuropsychology of Music. New York: Plenum Press, 1982, pp. 279–298.

Moorer, James A. "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer." Doctoral Dissertation, Department of Computer Science, Stanford University, 1975. Department of Music Report STAN-M-3.

Moorer, James A. "About this Reverberation Business." *Computer Music Journal* 3(2):13–28, 1979.

Morrill, Dexter. "The Dynamic Aspects of Trumpet Phrases." (French version: "Aspects dynamiques du Phrase de la Trompette," translated by Emmanuel Gresset). Paris: IRCAM, 1980.

Petersen, T. L. "Acoustic Signal Processing in the Context of a Perceptual Model." Ph. D. Dissertation, Computer Science Department, University of Utah, 1980.

Petit Larousse illustré. Paris: Librairie Larousse, 1979.

Piszczalski, M., and B. Galler. "Automatic Music Transcription." *Computer Music Journal* 1(4):24–31, 1977.

Piszczalski, M., B. Galler, R. Bossemeyer, and F. Looft. "Performed Music: Analysis, Synthesis, and Display by Computer." *Journal of the Audio Engineering Society* 29(1/2):38–46, 1981.

Plomp, Reinier. *Aspects of Tone Sensation*: A Psychophysical Study. New York: Academic Press, 1976.

Plomp, Reinier, and H. J. M. Steeneken. "Place Dependence of Timbre in Reverberant Sound Fields." *Acustica* 28(1):51-59, 1973.

Risset, Jean-Claude. "Computer Study of Trumpet Tones." Murray Hill, New Jersey: Bell Laboratories. Typewritten mss., 1966. *Journal of the Acoustical Society of America* 38:912, 1965 (abstract only).

Risset, Jean-Claude, and D. Wessel. "Exploration of Timbre by Analysis and Synthesis." In D. Deutsch, ed. *The Psychology of Music.* New York: Academic Press, 1982.

Saldanha, E. L, and John F. Corso. "Timbre Cues and the Identification of Musical Instruments." *Journal of the Acoustical Society of America* 36:2021–2026, 1964.

Schönberg, A. *HarmonieLehre.* Vienna: Universal, 1922. Translated as *Theory of Harmony* by Roy E. Carter. Berkeley, California: University of California Press, 1978.

Schubert, Earl D. Personal communication, 1982.

Der Sprach-Brockhaus. Wiesbaden: Brockhaus, 1970.

Strawn, John. *SYNTH Manual.* Stanford, California: CCRMA on-line documentation, 1979.

Strawn, John M. 1981. "Approximation and Syntactic Analysis of Amplitude and Frequeny Functions for Digital Sound Synthesis." *Computer Music Journal* 4(3):3–24.

Strawn, John. "eMerge — A Generalized Merge File Editor (Proposal)." 17 August 1982. Manuscript.

Wessel, David L. "Timbre Space as a Musical Control Structure." *Computer Music Journal* 3(2):45–52.

Winckel, Fritz. *Phänomene des musikalischen Hörens.* Berlin: Max Hesses Verlag, 1960. Translated (poorly) as *Music, Sound and Sensation: A Modern Exposition.* New York: Dover, 1967.

Captions for Figures

Figure 1. Schematic representation of the model suggested for certain aspects of auditory perception. The time-varying signal at the top is analyzed by the auditory system into a number of time-varying spectral components. Very early in the perceptual process, these spectral components are assigned to virtual sources. Thereafter, timbre, loudness, instrumental source, and other perceptual attributes are assigned to each virtual source.

Figure 2. Graphic representation of a musical passage. Is timbre explicitly assigned for all of the very short notes at "A" or "B"? See text for discussion.

Figure 3. Schematic representation of the phase vocoder. The original signal $x(t)$ is analyzed by the equivalent of a series of band-pass filters; each rectangle in the large vertical box at the left represents one such filter. Each filter produces a complex (real/imaginary) time-varying signal which can easily be converted into more intuitively appealing time-varying amplitude and frequency functions. Either of these representations can be used to resynthesize another signal $y(t)$, which will be identical to $x(t)$ as long as the real/imaginary (or the amplitude/frequency) functions remain unchanged. Figures 4–7 were generated using the phase vocoder.

Figure 4. Two clarinet tones (concert C-sharp 4, A3) as analyzed by the phase vocoder. For this recording, I told the clarinetist to play as legato as possible. The horizontal axis shows time, for a total duration of 2 sec. The phase vocoder was initialized so that one harmonic of the *second* tone would fall into one channel of the phase vocoder output; the channel center frequencies were thus approximately 220 Hz apart. Each thick horizontal bar represents one spectral component. The height of each bar gives the amplitude of the component in dB, with the tallest bar corresponding to the loudest component (0 dB). The frequency of the component is shown by the vertical position of the bar. The second harmonic of the first tone is essentially missing in this analysis.

Figure 5. Individual amplitude and frequency functions for the first ten channels of the tones shown in Figure 4. The frequency function for a given channel of the phase vocoder analysis is placed directly above the amplitude function for the same channel. Amplitude plots are on an arbitrary linear scale, which changes from channel to channel. The harmonic numbers for the two tones are written in as "H1," "H2," etc. The channel numbers are displayed between each amplitude/ frequency pair.

Figure 6. A closeup of time .92 sec through 1.0 sec from Figure 4. See text for discussion.

Figure 7. Individual amplitude and frequency functions, as in Figure 5, but with the time axis "zoomed in" to show time .8 – 1.1 sec. See text for discussion.

Figure 1:

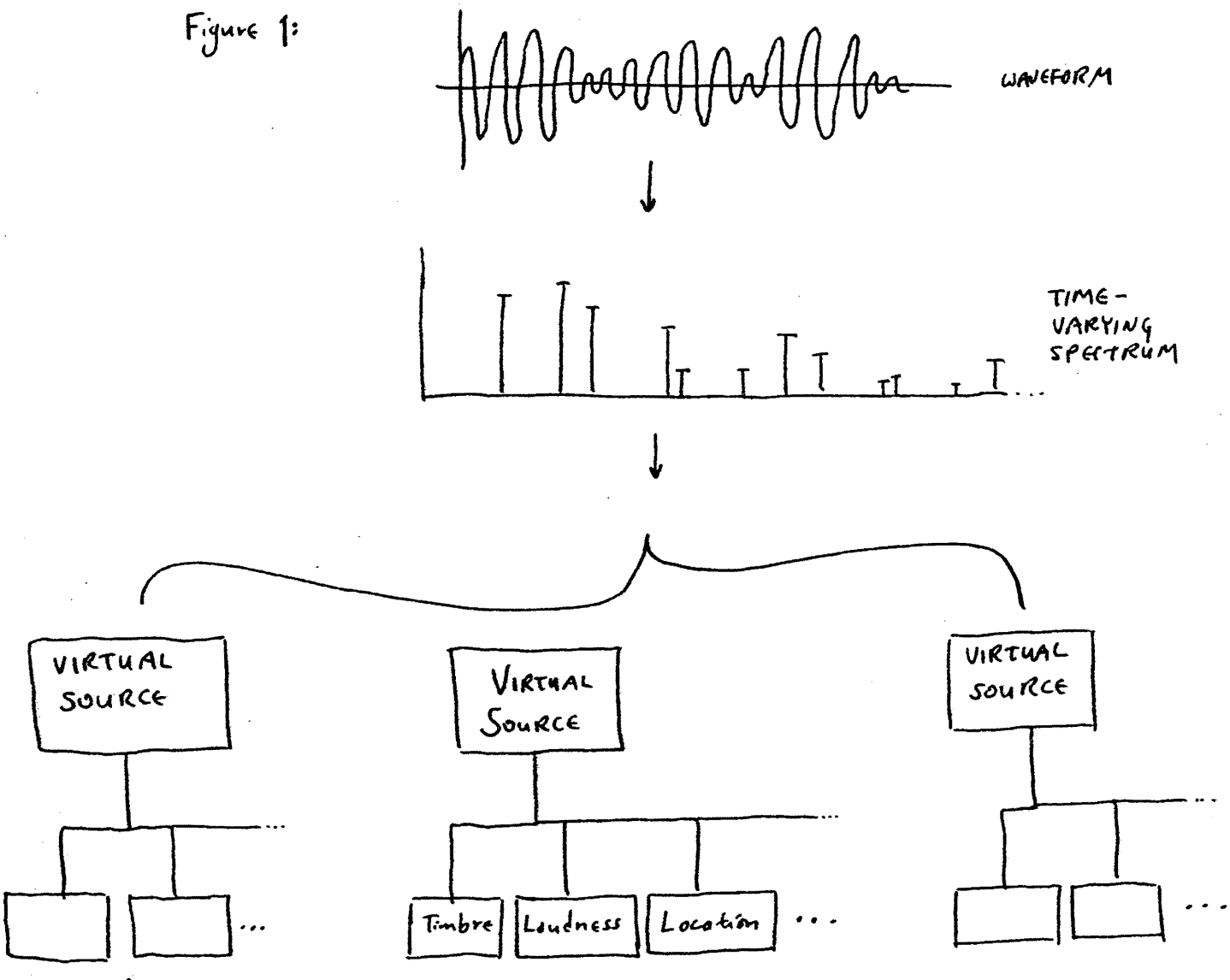
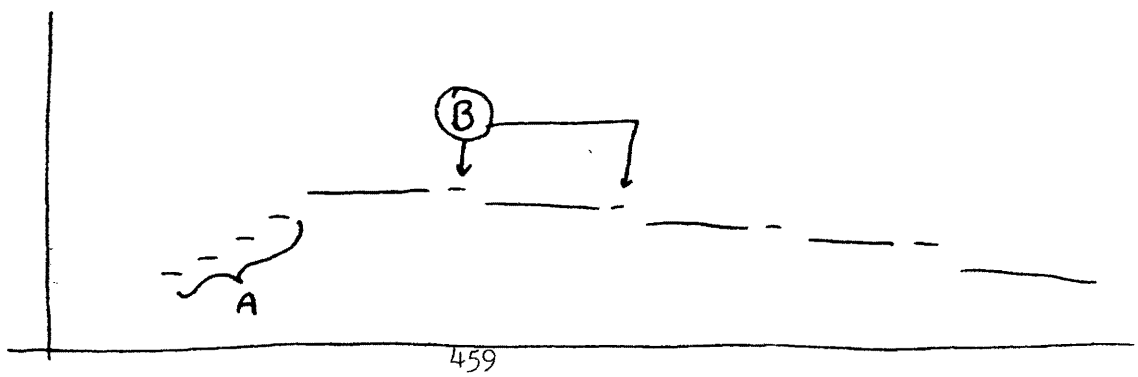


Figure 2:



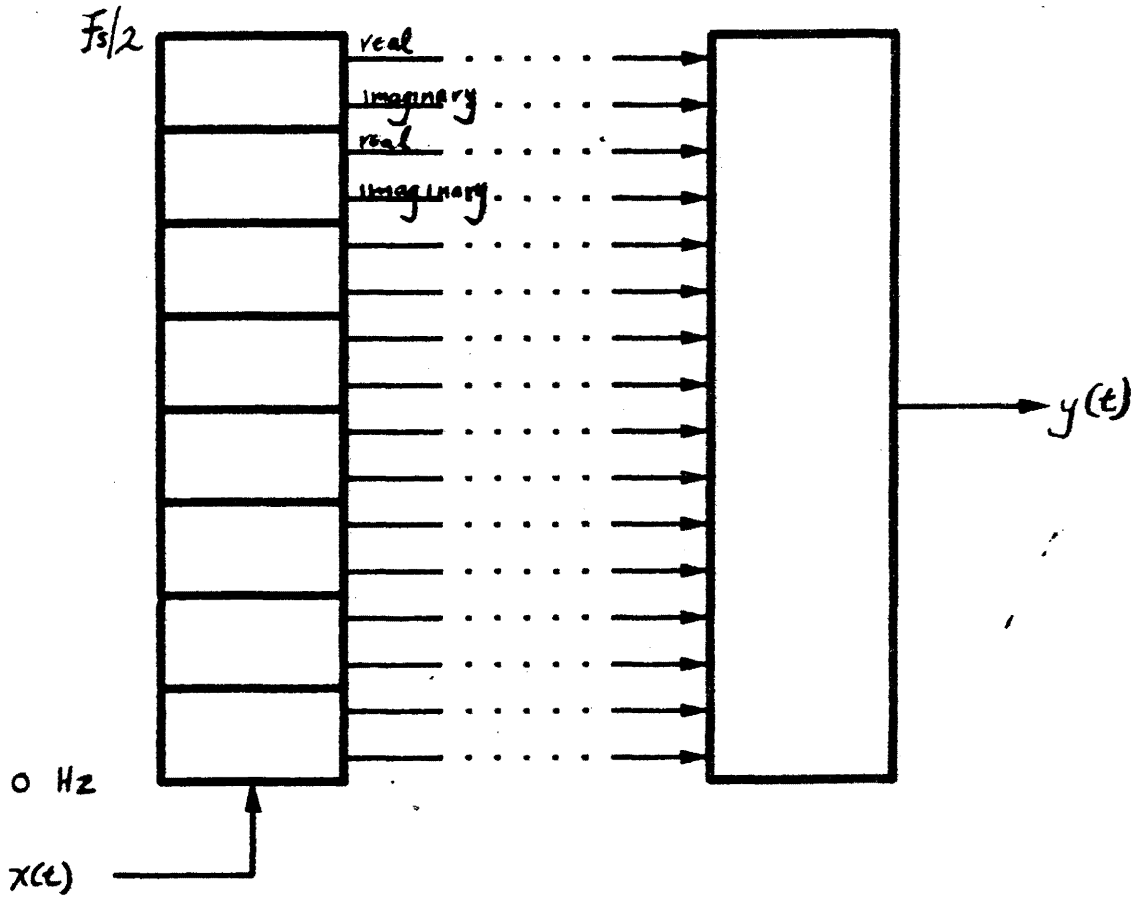
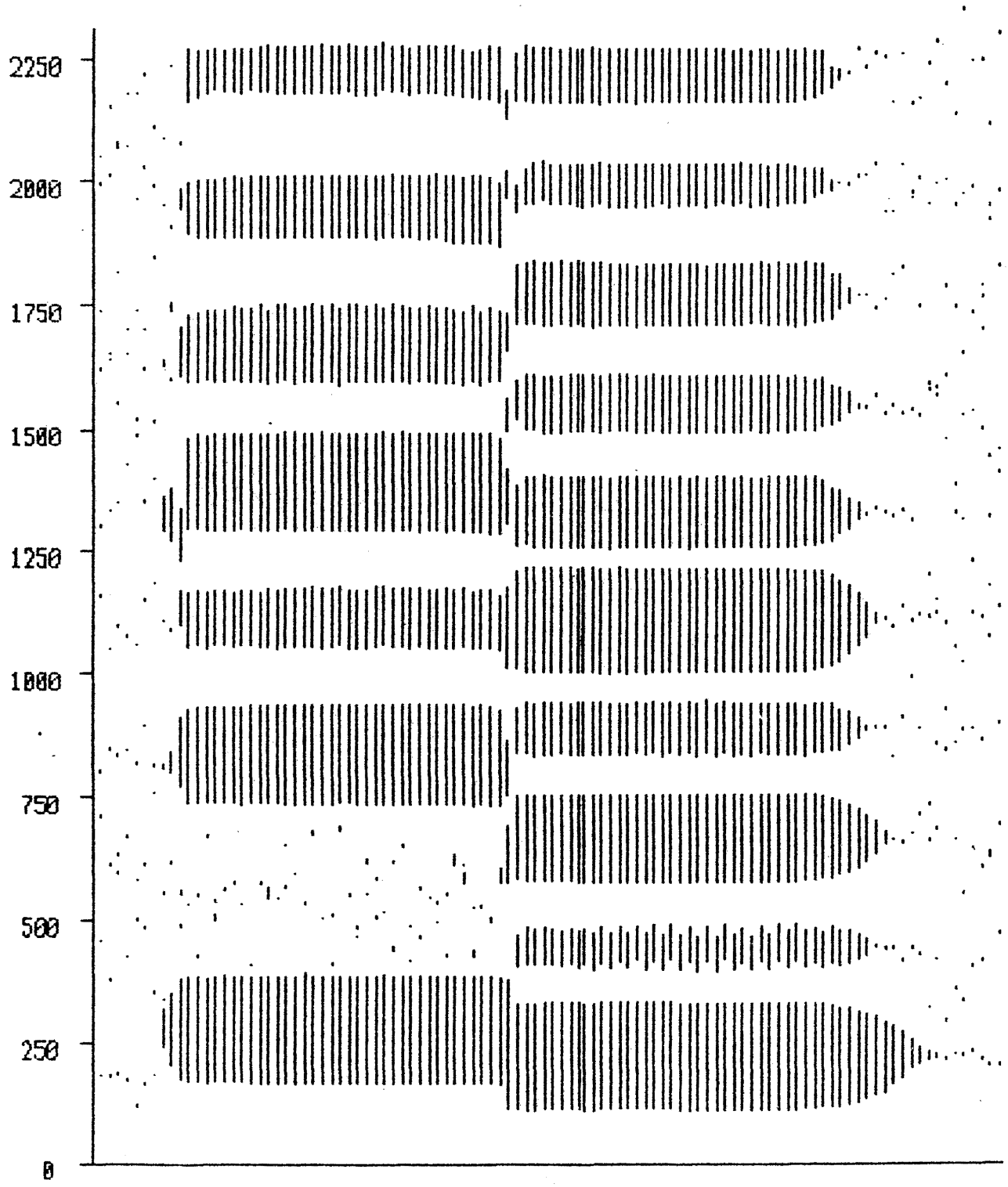


Figure 3

12 Sept 82

Figure 4



12 Sept '82 C044

Figure 5 p.1

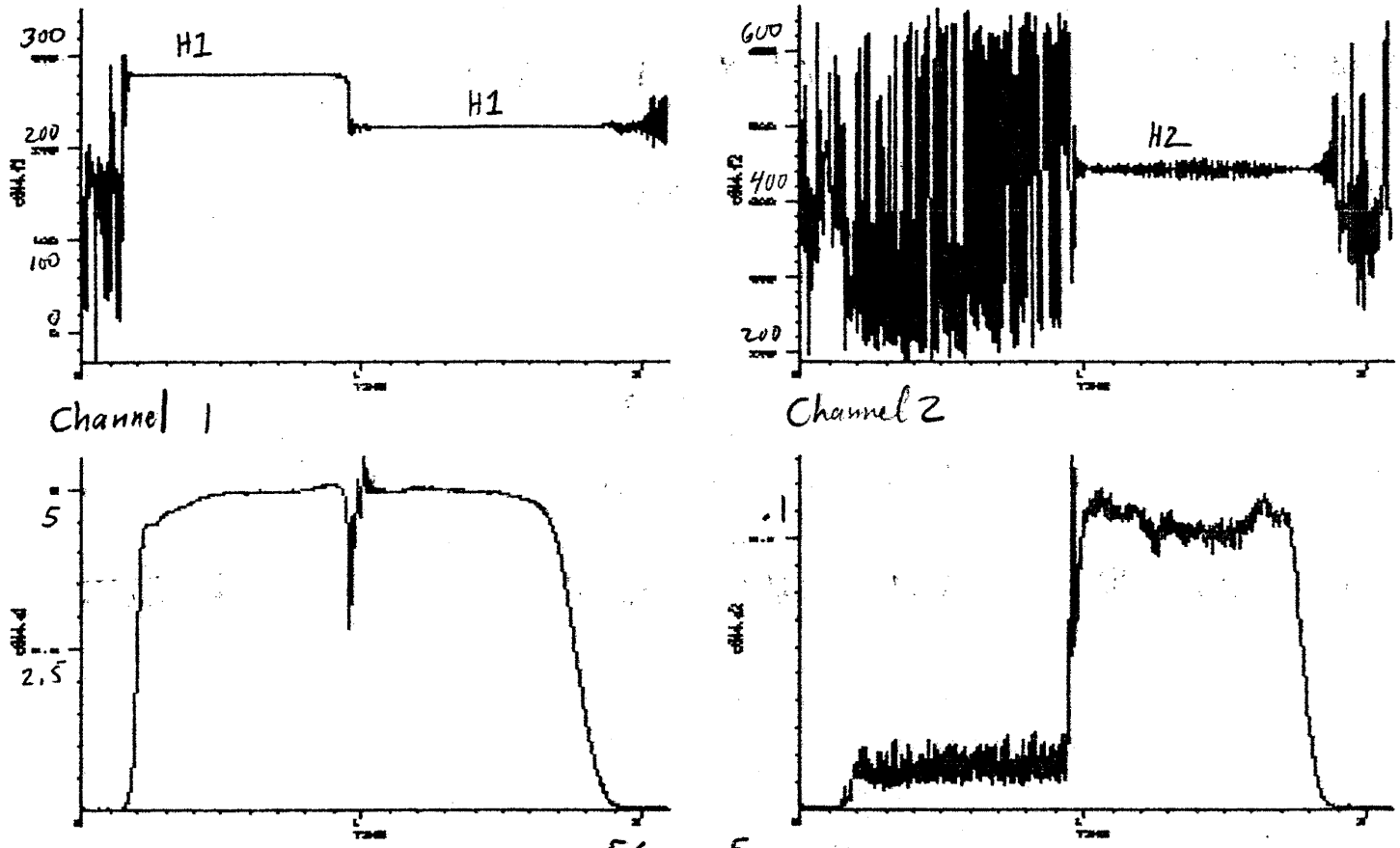
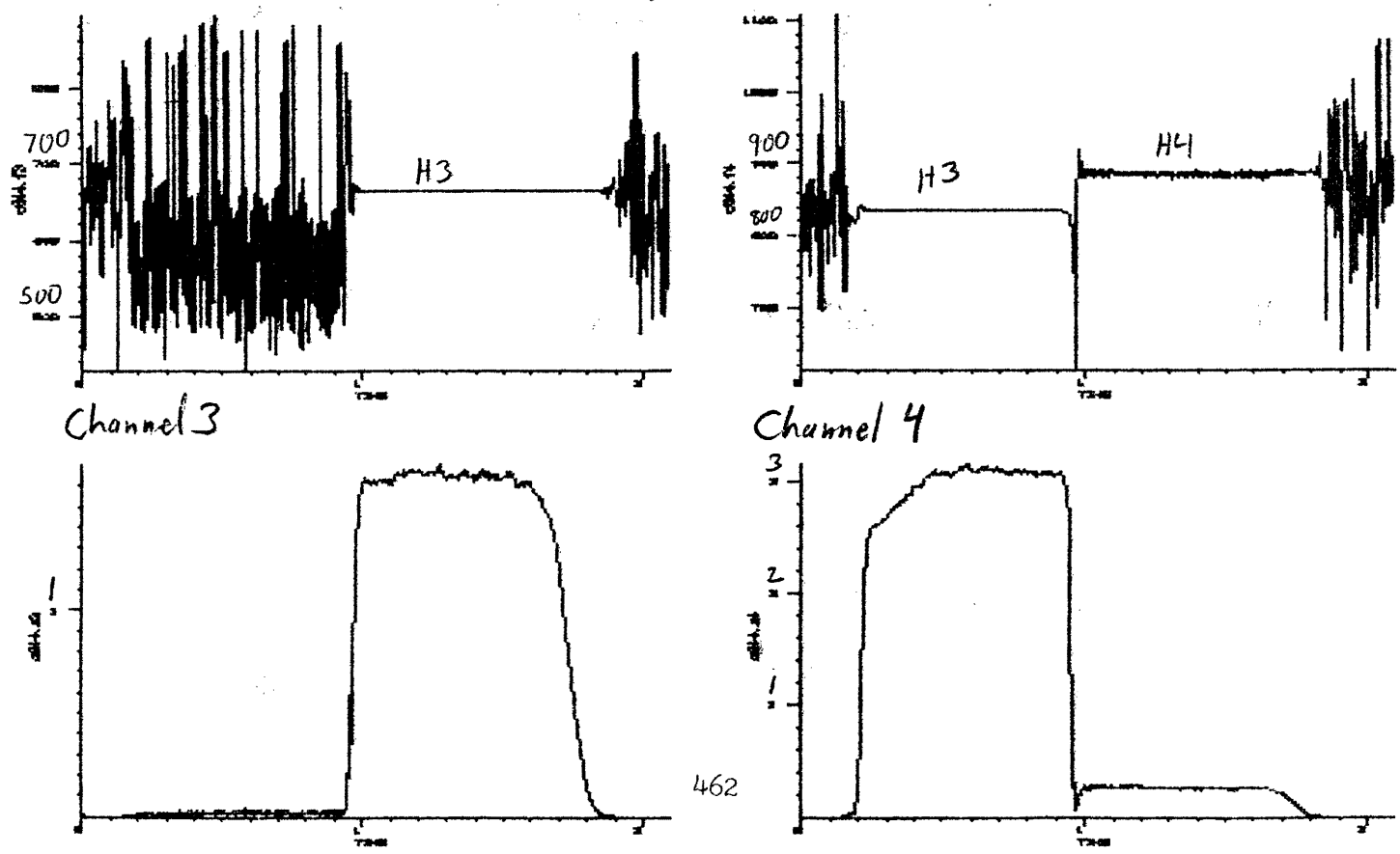
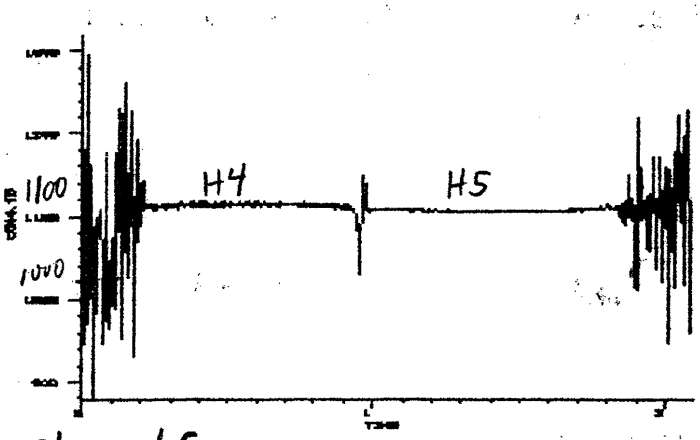
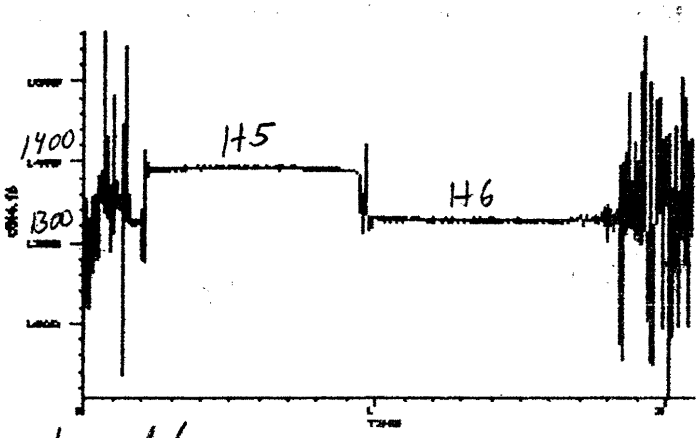


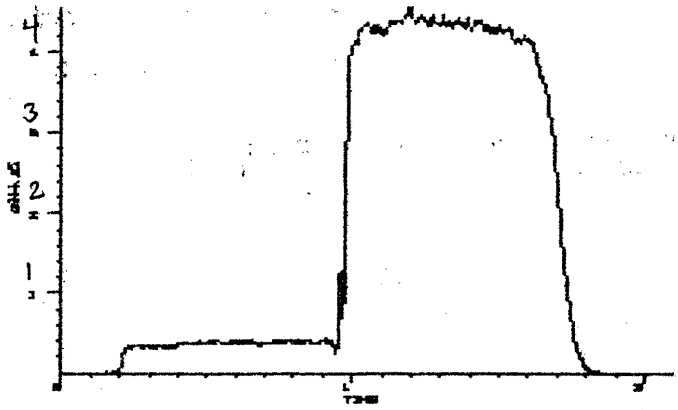
Figure 5







Channel 5



Channel 6

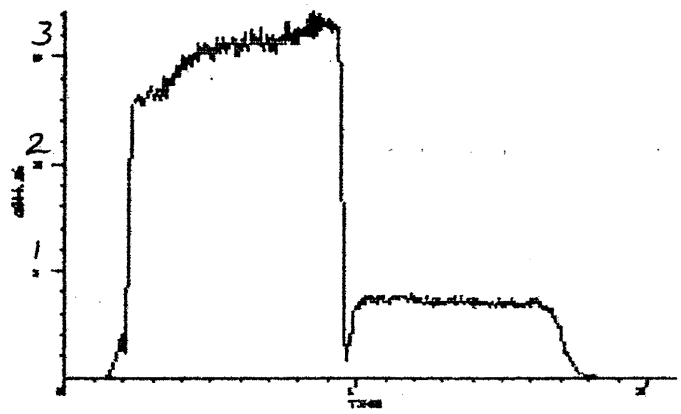
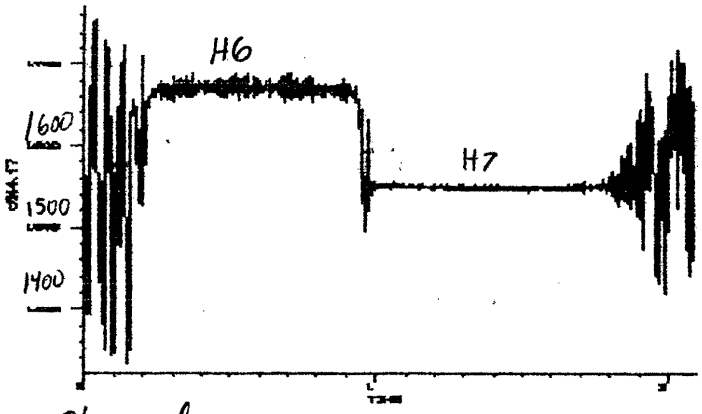
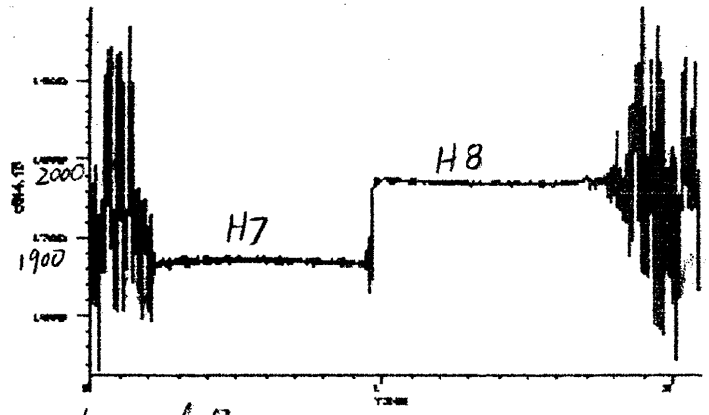
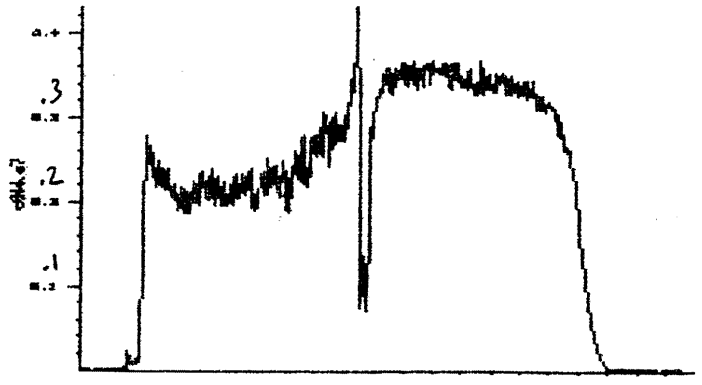


Figure 5

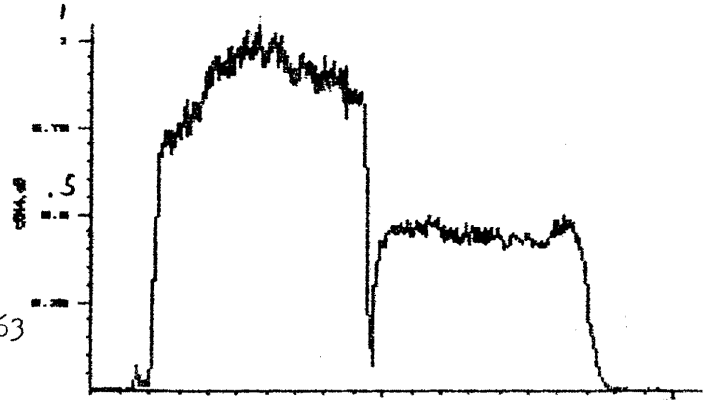


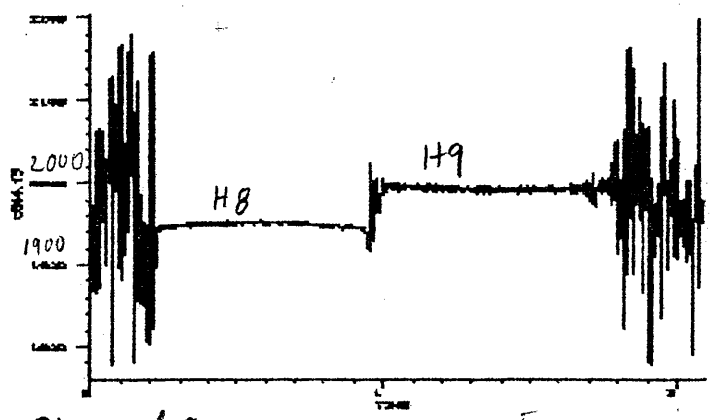


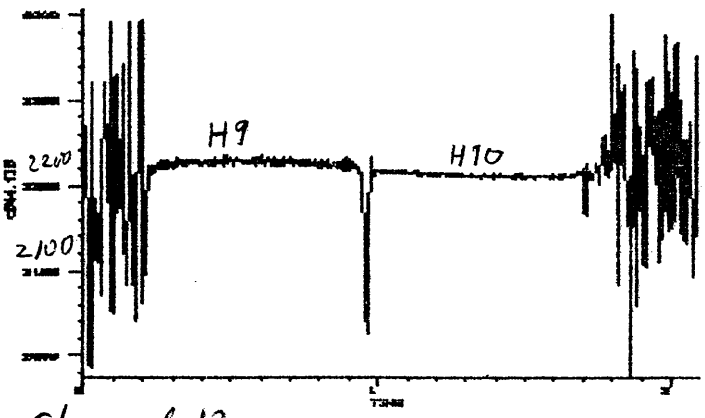
Channel 7



Channel 8

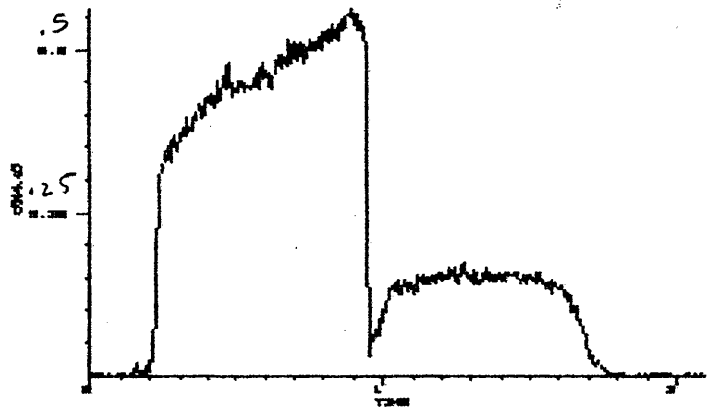






Channel 9

Channel 10



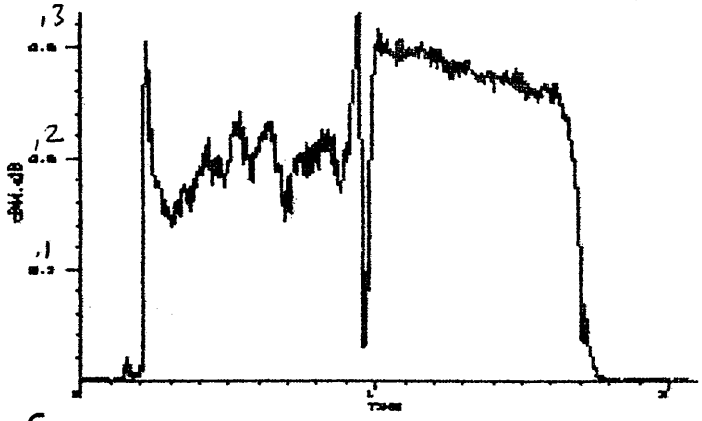
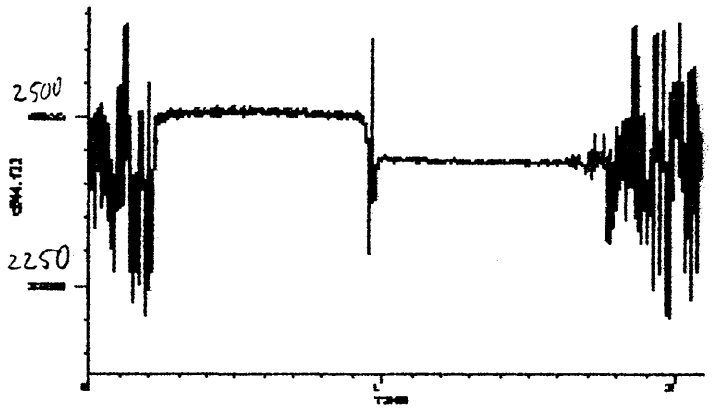
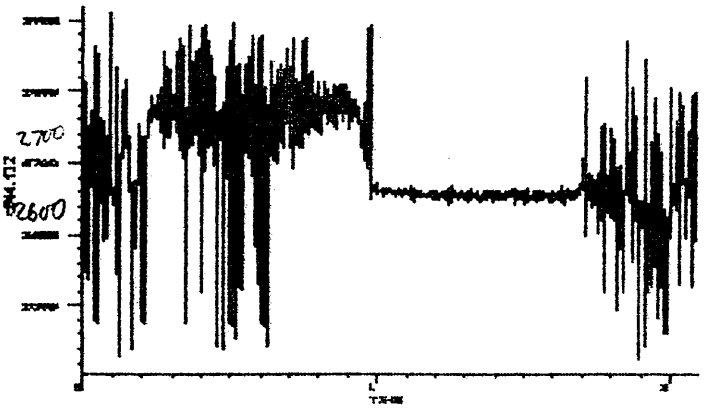
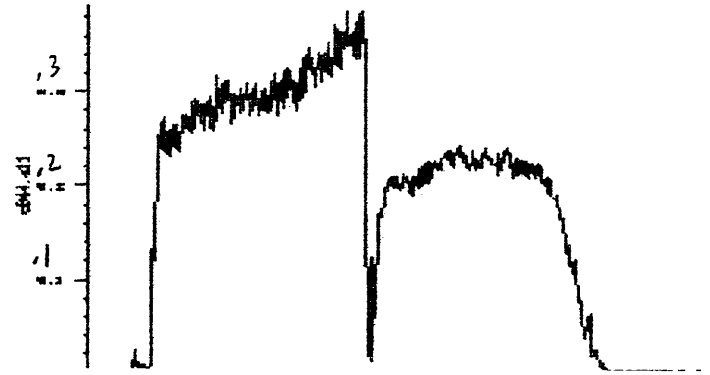


Figure 5







12 Sept 82

Figure 6